Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

SEQUENTIAL FIRST-ORDER RESPONSE SURFACE METHODOLOGY AUGMENTED WITH DIRECT GRADIENTS

Yunchuan Li

Department of Electrical and Computer Engineering
the Institute for Systems Research
University of Maryland
College Park, MD 20742, USA

Michael Fu

Robert H. Smith School of Business
the Institute for Systems Research
University of Maryland
College Park, MD 20742, USA

ABSTRACT

We introduce Direct Gradient Augmented Response Surface Methodology (DiGARSM), a new sequential first-order method for optimizing a stochastic function based on Response Surface Methodology (RSM). In this approach, gradients of the objective function with respect to the desired parameters are utilized in addition to response measurements. We establish convergence of the proposed method. We compare methods that use only response information and those that use only gradient information with the proposed approach, which uses both. Moreover, we conduct numerical simulations to illustrate the effectiveness of the proposed method.

1 INTRODUCTION

Response surface methodology (RSM) is a popular local metamodel method in simulation optimization, which sequentially determines variables that optimize the response of a system. RSM was first described in Box and Wilson (1951) and was applied to a real physical system. Since that time, the use of RSM has been extended successfully to many other scientific fields such as biometrics (Mead and Pike 1975), industrial engineering (Wu 1964) and materials science (Horng et al. 2008). RSM is often applied in optimizing stochastic simulation models. One of the earliest case studies is given in Keyzer et al. (1981). Other examples of RSM in simulation include Fu (1994) and Carson and Maria (1997). More recent developments of RSM in simulation are discussed in Barton and Meckesheimer (2006), Bartz-Beielstein and Preuss (2007) and Law et al. (2007). Myers et al. (1989) gives a review of developments in RSM in the period 1966-1988. More recent overviews of RSM can be found in Kleijnen (2015).

Traditionally, RSM views the system to be optimized as a black box and is able to obtain the input-output pairs (variable-response pairs) from the model. It uses a sequence of local experiments that leads to the optimum. In each local experiment, a number of input-output pairs are observed in a small region. A metamodel, which is usually a first or second-order polynomial model, is then used to fit the response surface. Steepest descent (or ascent) is performed to determine the next region to be explored, where the search direction is given by the fitted model. The fit and search process is repeated until a satisfactory result has been obtained; see Kleijnen (2015) for details. To determine input points to measure in each local experiment, several design methods are presented, e.g., factorial design, PlackettBurman design (Plackett and Burman 1946) and simplex design. More complex design methods include robust parameter design. A successful design should be examined based on several criteria, such as prediction variance (Box and Draper 1975). Experiment design and optimization method for multiple-response problems have also been studied. For example, Kim and Draper (1994) considers designs for systems with two correlated responses, and more general multiple-response systems are studied in Krafft and Schaefer (1992).

978-1-5386-6572-5/18/$31.00 ©2018 IEEE

Though RSM is only a heuristic (Kleijnen 2015), it works well in applications when relatively accurate response measurements are available. However, the measurement on the output of the system is often noisy, which could lead to unstable behavior of RSM. This is one leading motivation for our research: is there any additional information that we may utilize to cancel the negative impact of noisy response measurements?

In some simulation settings, direct gradient information may also be available, i.e., in addition to the performance measure of the system, the gradient of the performance measurement with respect to (w.r.t.) the parameters of interest, is also included in the output responses. A number of techniques to estimate the gradient of performance measure through samples have been proposed. Examples of such methods include perturbation analysis (Ho and Cao 1983; Glasserman 1991) and the likelihood ratio method (Rubinstein and Shapiro 1993). The gradient estimate has been applied extensively to stochastic approximation (Fu 1994). Therefore, when gradient measurements are also available, one way to potentially improve traditional RSM is to combine gradient information in fitting the metamodel. One big question here is how to incorporate the extra gradient measurements. A similar question is addressed in the context of stochastic approximation (Chau et al. 2014a; Chau et al. 2014b), where both gradient and response measurements are utilized with proven record of improvement. Therefore, we hope that the gradient measurements will help improve the performance of RSM.

With additional gradient information, a modified regression model-Direct Gradient Augmented Regression (DiGAR)-is investigated in Fu and Qu (2014). DiGAR fits a regression model using both response and gradient information with a least squares approach. This regression model shows great potential in the presence of significant response measurement noise. Under some mild assumptions, it is also shown that the estimator of the gradient is unbiased. Therefore, we expect the modified RSM with DiGAR model will perform better than traditional RSM with regular least-squares regression model. Moreover, since gradient augmented RSM uses both response and gradient measurements, we also believe that in cases where gradient information is unreliable but response measurement is accurate, i.e., high variance in gradient measurement but low in response measurement, the modified RSM should still perform well.

Another drawback in applying RSM is the lack of theoretical convergence guarantee. Some prior work related to the convergence property of RSM include stopping rules (Miró-Quesada and Del Castillo 2004) and confidence regions (Cahya et al. 2004). Theoretical performance of RSM incorporated with trust region method is presented by Chang et al. (2007). However, to the best of our knowledge, there is little research on the convergence analysis of RSM. Thus, we are also interested in establishing that RSM, including the version augmented with direct gradient information, converges to the optimal point of objective function.

In this paper, we propose an iterative local metamodel method called *Direct Gradient Augmented Response Surface Methodology* (DiGARSM). In each iteration, responses and their gradients w.r.t. to parameter of interest are observed in a local region. These responses are used to fit DiGAR model to determine the search direction. We examine the potential improvements in efficiency when the additional gradient information is present compared to traditional RSM and search model that uses only gradient information. We also investigate the performance of DiGARSM when the variance in gradient measurements is significant. Numerical experiments are carried out to illustrate the power of gradient information on the modified RSM model. Results show that DiGARSM is less sensitive to measurement variances compared to the one that uses only response or only gradient information: it adjusts the search direction with the additional measurements. Moreover, simulations also show that DiGARSM leads to faster convergence than traditional RSM. In summary, contributions of this work include:

1. We propose a modified RSM that incorporates both response and gradient measurements.
2. We provide preliminary convergence analysis for both RSM and DiGARSM.
3. We provide numerical experiments that illustrate the efficiency and robustness against measurement variances of our algorithm.

The rest of the paper is organized as follows: Section 2 describes the underlying problem and proposes DiGARSM, as well as models that use only response or gradient measurements, for one-dimensional problems. Section 3 extends the investigation to multidimensional problems and proposes search directions under such settings. Section 4 presents some preliminary theoretical results. Numerical examples for two different objective functions are shown in Section 5 to illustrate the potential power of this new method. Finally, we conclude and provide some future research directions in Section 6.

2 ONE-DIMENSIONAL PROBLEMS

Consider a stochastic optimization problem

$$\min_{x\in R} f(x) = \min_{x\in R} \mathbb{E}[\tilde{f}(x)],$$

where $f:\mathbb{R}\to\mathbb{R}$, $f(x)=\mathbb{E}[\tilde{f}(x)]$, and $\tilde{f}(x)$ is a noisy sample of response at x. We assume that $f(x)$ is a convex function but its closed form is not available to us. In addition to access to noisy response samples, we can also obtain noisy direct gradient $\nabla\tilde{f}(x)$ estimates at the same time. Let $\nabla f(x)=\mathbb{E}[\nabla\tilde{f}(x)]$. Further assume homogeneous noise in the estimation, let $\tilde{f}(x)=f(x)+\varepsilon_x$ and $\nabla\tilde{f}(x)=\nabla f(x)+\delta_x$, where $\varepsilon_x\sim N(0,\sigma_f^2)$, $\delta_x\sim N(0,\sigma_g^2)$, $\mathrm{Cov}(\varepsilon_x,\delta_y)=\rho$ for $x=y$ and 0 otherwise.

RSM generates a sequence of iterates $\{x_k\}$. At iteration k, input-output pair samples around x_k are taken and fitted to a linear model. The next iterate is found using the recursion

$$x_{k+1} = x_k - a_k g_k, \tag{1}$$

where g_k is the derivative, which is obtained from the fitted linear model, and $a_k>0$ is the step size. Suppose that at iteration k we sample symmetrically around x_k at x_k+c_k, x_k-c_k for n times each, where c_k is some positive number that changes at each iteration. Denote the optimal point by

$$x^* = \arg\min_x f(x).$$

We hope that recursive runs of Equation (1) will lead to x^*. To make our notation cleaner, we denote x_k^i as the i-th sampled set of points in the $k-$th iteration, i.e., $x_k^i=\{x_{k,-}^i,x_{k,+}^i\}=\{x_k-c_k,x_k+c_k\}$ for all i. With a little abuse of notation, let $\tilde{f}(x_k^i)=\{\tilde{f}(x_{k,-}^i),\tilde{f}(x_{k,+}^i)\}$ and $\nabla\tilde{f}(x_k^i)=\{\nabla\tilde{f}(x_{k,-}^i),\nabla\tilde{f}(x_{k,+}^i)\}$ be the set of response and gradient estimates of the set x_k^i. To simplify equations, define set $\mathscr{S}=\{+,-\}$.

Now assume that the algorithm has proceeded to iteration k. In the rest of this section, we will discuss the metamodel to be fit and the corresponding gradient estimates under different assumptions.

Remark 1 Traditionally, RSM fits samples to a second-order polynomial in the last step. In this paper we will focus on fitting samples to linear models to obtain the preliminary theoretical results in Section 4.

2.1 Direction by Ordinary Linear Regression (RSM)

First let us assume that we can only access the responses at x_k^i, i.e., we have $\tilde{f}(x_k^i)$ for $i=1,2,\ldots,n$. To estimate the derivative g_k, we fit the $2n$ data to an ordinary linear regression model

$$\hat{f}(x) = \beta_{k0} + \beta_k x.$$

The least squares approach minimizes the sum of squared residuals given by

$$\begin{aligned} L &= \sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(\tilde{f}(x_{k,j}^i)-\hat{f}(x_{k,j}^i))^2 \\ &= \sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(\tilde{f}(x_{k,j}^i)-\beta_{k0}-\beta_k x_{k,j}^i)^2. \end{aligned}$$

The optimal β_k that minimizes L is given by

$$\hat{\beta}_k = \frac{(1/2n)\sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(x_{k,j}^i - \bar{x}_k)(\tilde{f}(x_{k,j}^i) - \bar{\tilde{f}}_k)}{(1/2n)\sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(x_{k,j}^i - \bar{x}_k)^2}, \tag{2}$$

where $\bar{x}_k$ and $\bar{\tilde{f}}_k$ are the means over all sampled points and sample responses at iteration k, respectively. The derivative is then given by $\hat{\beta}_k$, i.e., $g_k = \hat{\beta}_k$.

2.2 Direction with Gradient Estimates

In this part, we consider systems where only direct gradient estimates are used, i.e., for x_k^i, we use only $\nabla\tilde{f}(x_k^i)$ and not $\tilde{f}(x_k^i)$, as in the Gradient Surface Methodology (GSM) of Ho et al. (1992). Then one suitable gradient estimate can be calculated by fitting gradients

$$\nabla\hat{f}(x) = \beta_k, \forall i = 1,2,\ldots,2n.$$

The β_k that minimizes the sum-of-squared-error becomes the optimal estimate in the least squares sense. Therefore, β_k is given by minimizing

$$\begin{aligned} L &= \sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(\nabla\tilde{f}(x_{k,j}^i) - \nabla\hat{f}(x_{k,j}^i))^2 \\ &= \sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(\nabla\tilde{f}(x_{k,j}^i) - \beta_k)^2. \end{aligned}$$

The minimizer (and hence g_k) is the mean over all samples, i.e.,

$$g_k = \hat{\beta}_k = \frac{1}{2n}\sum_{j\in\mathscr{S}}\sum_{i=1}^{n}\nabla\tilde{f}(x_{k,j}^i) = \nabla\bar{\tilde{f}}_k. \tag{3}$$

2.3 Direction by Augmented Regression with Gradient

Now we present DiGARSM, in which we assume both response measurement and direct gradient estimates are available at the time of sampling, that is, we can acquire both $\tilde{f}(x_k^i)$ and $\nabla\tilde{f}(x_k^i)$ when we sample point x_k^i. Then we have $2n$ noisy estimates of the responses and gradients. To obtain a proper gradient estimate from these samples, DiGARSM first fit these data to least square $\alpha-$DiGAR model proposed in Fu and Qu (2014)

$$\begin{aligned} \hat{f}(x) &= \beta_{k0} + \beta_k x, \\ \nabla\hat{f}(x) &= \beta_k, \end{aligned}$$

by minimizing the weighted sum of loss

$$\begin{aligned} L &= \alpha\sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(\tilde{f}(x_{k,j}^i) - \hat{f}(x_{k,j}^i))^2 + (1-\alpha)\sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(\nabla\tilde{f}(x_{k,j}^i) - \nabla\hat{f}(x_{k,j}^i))^2 \\ &= \alpha\sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(\tilde{f}(x_{k,j}^i) - c_k - \beta_k x_{k,j}^i)^2 + (1-\alpha)\sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(\nabla\tilde{f}(x_{k,j}^i) - \beta_k)^2, \end{aligned}$$

where $\alpha \in [0,1]$ is the weight on response measurements and needs to be specified by users.

The minimizing β_k (and hence g_k) is given by

$$g_k = \hat{\beta}_k = \frac{\alpha(1/2n)\sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(x_{k,j}^i - \bar{x}_k)(\tilde{f}(x_{k,j}^i) - \bar{\tilde{f}}_k) + (1-\alpha)\nabla\bar{\tilde{f}}_k}{(1/2n)\alpha\sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(x_{k,j}^i - \bar{x}_k)^2 + (1-\alpha)}. \tag{4}$$

The for DiGARSM takes advantage of both response and gradient samples. Therefore, we expect it will outperform RSM, which only fits response or gradient samples.

Note that if we set $\alpha = 1$, the gradient information is not utilized, and Equation (4) becomes Equation (2). If $\alpha = 0$, only gradient information is used, and thus reduces to Equation (3), which resembles the Kiefer-Wolfowitz Stochastic Approximation (SA) algorithm, except it uses the sample mean of the gradient estimates.

3 MULTIDIMENSIONAL PROBLEMS

In this section we propose the DiGARSM algorithm for problems with more than one dimension, i.e., our objective is to solve

$$\min_{x\in R^d} f(x) = \min_{x\in R^d} \mathbb{E}[\tilde{f}(x)],$$

where $d > 1$ is the number of dimensions. Similar to one-dimensional cases, assume $\tilde{f}(x) = f(x) + \varepsilon_x$ and $\nabla\tilde{f}(x) = \nabla f(x) + \Delta_x$, where $\Delta_x \sim N(0, \delta_x I)$. At iteration k, we apply a full factorial design to determine the points to sample with a total $2^d n$ of samples. The $i-$th set of sample points $x_k^i = \{x_{k,j}^i | j \in \mathscr{S}^d\}$, the set of response measurements $\tilde{f}(x_k^i)$ and the set of gradient measurements $\nabla\tilde{f}(x_k^i)$ are defined analogously. The recursive formula is the same as Equation (1) in one-dimensional problems, except that the iterate x_k and gradient g_k will be of d dimensions. Additionally, the regression models for the three different designs in Section 2 remain unchanged, although the solution may seem more complicated.

3.1 Direction by Ordinary Linear Regression in Multidimensional Problem

With only responses available, RSM fits

$$\hat{f}(x) = \beta_{k0} + \beta_k^T x,$$

where $\beta_k = [\beta_{k1}, \ldots, \beta_{kd}]^T$, by minimizing

$$\begin{aligned} L &= \sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n}(\tilde{f}(x_{k,j}^i) - \hat{f}(x_{k,j}^i))^2 \\ &= \sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n}(\tilde{f}(x_{k,j}^i) - \beta_{k0} - \beta_k^T x_{k,j}^i)^2. \end{aligned}$$

The optimal β_k is given by

$$\hat{\beta}_k = [\sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n}(x_{k,j}^i - \bar{x}_k)(x_{k,j}^i - \bar{x}_k)^T]^{-1}[\sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n}(x_{k,j}^i - \bar{x}_k)(\tilde{f}(x_{k,j}^i) - \bar{\tilde{f}}_k)]. \tag{5}$$

Consequently, the gradient estimate $g_k = \beta_k$.

3.2 Direction with Gradient Estimates in Multidimensional Problem

When only gradient information is utilized in multidimensional problems, we fit our data to a model similar to that in the one-dimensional case:

$$\nabla\hat{f}(x) = \beta_k.$$

By minimizing the sum of squared 2-norm of the error

$$L = \sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n} ||\nabla\tilde{f}(x_{k,j}^i) - \nabla\hat{f}(x_{k,j}^i)||^2,$$

we obtain the optimal β_k (and hence g_k)

$$g_k = \hat{\beta}_k = \frac{1}{2^d n}\sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n}\nabla\tilde{f}(x_{k,j}^i) = \nabla\bar{\bar{\tilde{f}}}_k, \tag{6}$$

which is also the mean over all sampled data.

3.3 Direction by Augmented Regression with Gradient in Multidimensional Problem

Similar to the one-dimensional case, DiGARSM assumes that noisy response and gradient information can be obtained, and the derivative estimate is found by fitting the α-DiGAR model for multidimensional problems

$$\hat{f}(x) = \beta_{k0} + \beta_k^T x,$$
$$\nabla\hat{f}(x) = \beta_k = [\beta_{k1}, \beta_{k2}, \dots, \beta_{kd}]^T.$$

The resulting gradient is found by minimizing

$$\begin{aligned} L &= \alpha_0\sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n}(\tilde{f}(x_{k,j}^i) - \hat{f}(x_{k,j}^i))^2 + \sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n}(\nabla\tilde{f}(x_{k,j}^i) - \nabla\hat{f}(x_{k,j}^i))^T W(\nabla\tilde{f}(x_{k,j}^i) - \nabla\hat{f}(x_{k,j}^i)) \\ &= \alpha_0\sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n}(\tilde{f}(x_{k,j}^i) - \beta_{k0} - \beta_k^T x)^2 + \sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n}(\nabla\tilde{f}(x_{k,j}^i) - \beta_k)^T W(\nabla\tilde{f}(x_{k,j}^i) - \beta_k), \end{aligned}$$

where $W = diag(\alpha_1, \dots, \alpha_d)$ is the weight matrix with $\sum_{i=0}^{d}\alpha_i = 1$, which also needs to be determined by practitioners.

The minimizing β_k (hence g_k) is given by

$$\hat{\beta}_k = [\alpha_0\sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n}(x_{k,j}^i - \bar{x}_k)(x_{k,j}^i - \bar{x}_k)^T + 2^d nW]^{-1}[\alpha_0\sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n}(x_{k,j}^i - \bar{x}_k)(\tilde{f}(x_{k,j}^i) - \bar{\bar{\tilde{f}}}_k) + 2^d nW\nabla\bar{\bar{\tilde{f}}}_k]. \tag{7}$$

Similar to the one-dimensional case, when $\alpha_0 = 0$, only gradient information is utilized, and when $\alpha_0 = 1$ only responses of the function are used, the algorithm becomes Kiefer-Wolfowitz (KW) stochastic approximation.

Remark 2 In order to illustrate the differences between the derivatives in the three approaches, and show that Equation (4) can be seen as a hybrid of Equations (2) and (3) (Equations (7), (5), (6), respectively for multidimensional problems), we provide our result in explicit form. In practice, it may be more convenient to write the solution in matrix form; see Fu and Qu (2014) for more details.

4 THEORETICAL RESULTS

We have the following results regarding traditional RSM and DiGARSM, respectively.

Proposition 1. *(Sequential First-Order RSM) Let $\{x_k\}$ be a sequence following recursion (1) and $g_k = \beta_k$ given in Equation* (2) *(one dimensional) or Equation* (5) *(multidimensional). If the following conditions hold:*

1. $\sum_{k=1}^{\infty} a_k = \infty$, $\sum_{k=1}^{\infty} a_k c_k < \infty$, $\sum_{k=1}^{\infty} a_k^2 < \infty$, *and* $\sum_{k=1}^{\infty} a_k^2 c_k^2 < \infty$ *for positive sequences* $\{a_k\}$ *and* $\{c_k\}$.
2. $\exists K_0,\ K_1 > 0$ *such that* $K_0||x-x^*|| \le ||\nabla f(x)|| \le K_1||x-x^*||$ *for all* $x \in R$.
3. $\nabla f(x)^T(x-x^*) > 0$ *for all* $x \neq x^*$ *(*f *is quasiconvex).*
4. $\exists B$ *such that* $\sigma_f^2 < B < \infty$ *and* $\sigma_g^2 < B < \infty$.
5. ε_x *and* δ_y *are independent for all* $x \neq y$, $\mathrm{Cov}(\varepsilon_x, \delta_y) = \rho$ *for all* $x \in R$ *and* $\mathrm{Cov}(\varepsilon_x, \delta_y) = 0$ *for* $x \neq y$.

Then x_k *converges to* x^* *in mean square.*

The proof follows closely the proof of Proposition 2, which will be addressed in our future research.

Proposition 2. *(Sequential First-Order DiGARSM) Let* $\{x_k\}$ *be a sequence following recursion* (1) *and* $g_k = \beta_k$ *given in Equation* (4) *(one-dimensional) or Equation* (7) *(multidimensional). Under the same assumptions as those in Proposition 1,* x_k *converges to* x^* *in mean square.*

To summarize, the algorithmic description of DiGARSM is shown in Algorithm 1.

Algorithm 1: Algorithmic description of DiGARSM.

```
Input: Initial point x_0, weights α_0, ..., α_d, positive sequences a_k and c_k
Output: Optimal point x* = arg min_x f(x)
k ← 0;
while stopping rule not met do
    Calculate the set of points to be sampled
    x_k^i for i = 1, 2, ..., n;
    Obtain samples of response and gradient:
    f(x_k^i) and ∇f(x_k^i) for i = 1, 2, ..., n;
    Calculate derivative estimate:
```

$$g_k = \begin{cases} \frac{\alpha(1/2n)\sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(x_{k,j}^i-\bar{x}_k)(\tilde{f}(x_{k,j}^i)-\bar{\tilde{f}}_k)+(1-\alpha)\nabla\bar{\tilde{f}}_k}{(1/2n)\alpha\sum_{j\in\mathscr{S}}\sum_{i=1}^{n}(x_{k,j}^i-\bar{x}_k)^2+(1-\alpha)} & (one-dimensional) \\ [\alpha_0\sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n}(x_{k,j}^i-\bar{x}_k)(x_{k,j}^i-\bar{x}_k)^T+2^d nW]^{-1}\cdot & \\ [\alpha_0\sum_{j\in\mathscr{S}^d}\sum_{i=1}^{n}(x_{k,j}^i-\bar{x}_k)(\tilde{f}(x_{k,j}^i)-\bar{\tilde{f}}_k)+2^d nW\nabla\bar{\tilde{f}}_k] & (multidimensional) \end{cases}$$

```
    x_{k+1} ← x_k − a_k g_k;
    k ← k + 1;
end
return x_k;
```

5 NUMERICAL SIMULATIONS

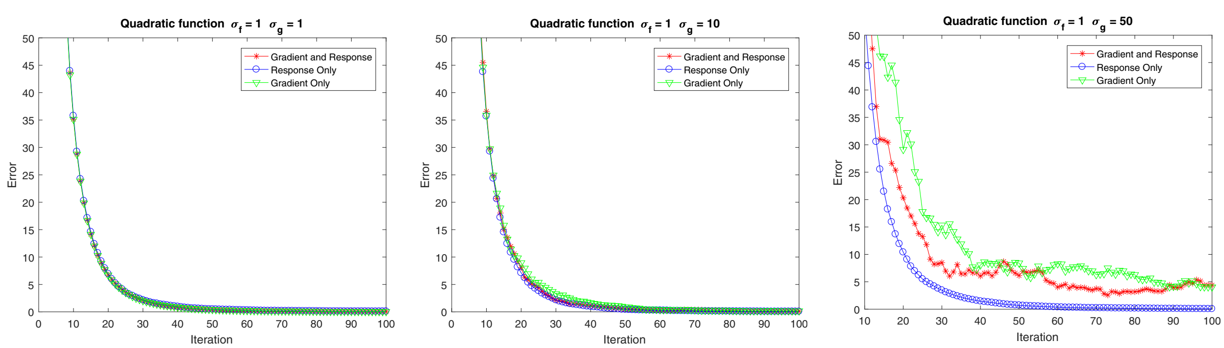
In this section we compare three different methods for gradient estimates: use both response and gradient information (DiGARSM), use only responses (RSM) and use only gradient information (KW stochastic search). These three designs will be referred to as designs 1, 2, and 3, respectively, henceforth. Two test functions, i.e., quadratic function and Booth function (Jamil and Yang 2013), are used to examine the efficiency of our designs, with the dimension (d) set to 2. The measurements of responses and gradients of the function at point x are corrupted with measurement noises ε_x and δ_x, respectively with $\varepsilon_x \sim N(0,\sigma_f^2)$ and $\delta_x \sim N(0,\sigma_g^2)$. The variances of the noises are taken from the set $\{1,10,50\}$. The initial point is set randomly and the same for all three methods. To compare the efficiency of three designs, the error of the function value of each iterate to the optimum ($f(x_k) - f(x^*)$) is calculated at each iteration. We run each simulation independently for 10 times and the average error is plotted. The weights (α) are set to be uniform, i.e., $\alpha_i = 1/3$ for $i = 0,1,2$. The positive sequences $\{a_k\}$ and $\{c_k\}$ are set to $a_k = 1/(10+k)$ and $c_k = (100+k)^{-\frac{1}{3}}$.

5.1 Quadratic Function

The quadratic function is given by $f(x) = x^T x$. Clearly, the global minimum is achieved at $x^* = (0,0)^T$ with $f(x^*) = 0$. The average error with different combinations of noises variances are shown in Figures 1 to 3. The y-axis is truncated to [0,50] to show details. Note that in Figure 3, the plot of Design 2 is out of the scope due to large error. When the variance of responses measurement (σ_f^2) and variance of gradient measurement (σ_g^2) are low (Figures 1a, 1b, 2a and 2b), all three algorithms perform similarly in error. When σ_f^2 is significantly larger (Figures 3a and 3b), the algorithm using only response information has inferior performances, while the other two designs applying gradient information still generate stable results. This makes sense, because RSM is only utilizing very noisy responses of the function but the other two use the less corrupt gradient. When σ_g^2 is relatively larger (Figures 1c and 2c), design 3 is affected most, especially for the first few iterations, whereas design 2 is unaffected, because it does not use gradient information at all. Design 1, where both responses and gradients are applied, is affected but to a lesser extent: its error increases for some iterations, but is still very close to that of design 2. When both responses measurements and gradient measurements are noisy (Figure 3c), all three designs are influenced compared with low noise variances. Design 2 becomes very unstable, like those in Figures 3a and 3b. As for design 1 and 3, although the gradient measurements are quite noisy, the two architectures still converge to the optimum.

When the gradient variance is large and the response measurement variance is reasonably small (e.g., Figures 1c and 2c), design 1 has a faster and more stable performance within the first 30 iterations, and after that the two designs perform similarly. We believe that the usage of more informative responses in DiGARSM leads to this result in the first few iterations. When the algorithm has reached a point where the error is small (i.e., after 30 iterations in our example), the variances of the measurements become comparable to the function values (because of the homogeneous noise assumption), which makes design 1 and design 3 perform similarly.

From Figures 1c, 2c and 3a to 3c, it is clear that the use of both gradients and responses of the objective function indeed improves the performance of RSM.

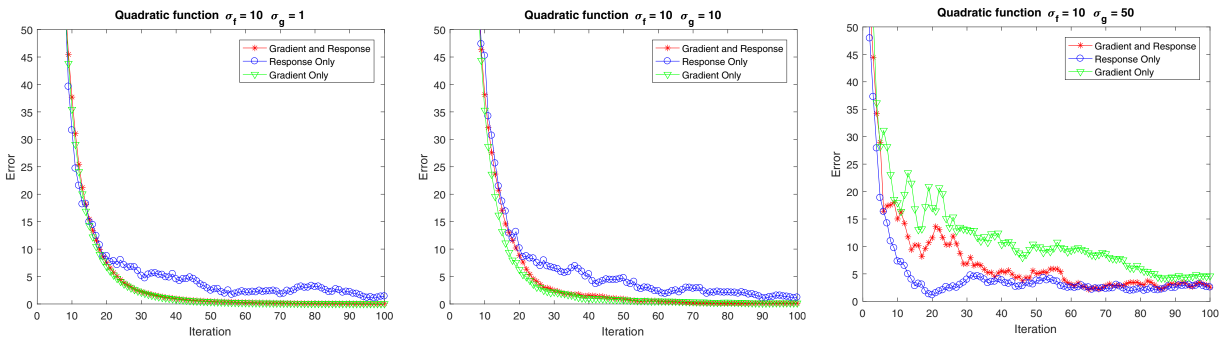


(a) $\sigma_g^2 = 1$. (b) $\sigma_g^2 = 10$. (c) $\sigma_g^2 = 50$.

Figure 1: Quadratic function: $\sigma_f^2 = 1$.

5.2 Booth Function

The Booth function is given by $f(x) = (x_1 + 2x_2 - 7)^2 + (2x_1 + x_2 - 5)^2$. The global minimum is achieved at $x^* = (1,3)$ with $f(x^*) = 0$. The average error with different combinations of noise variances are shown

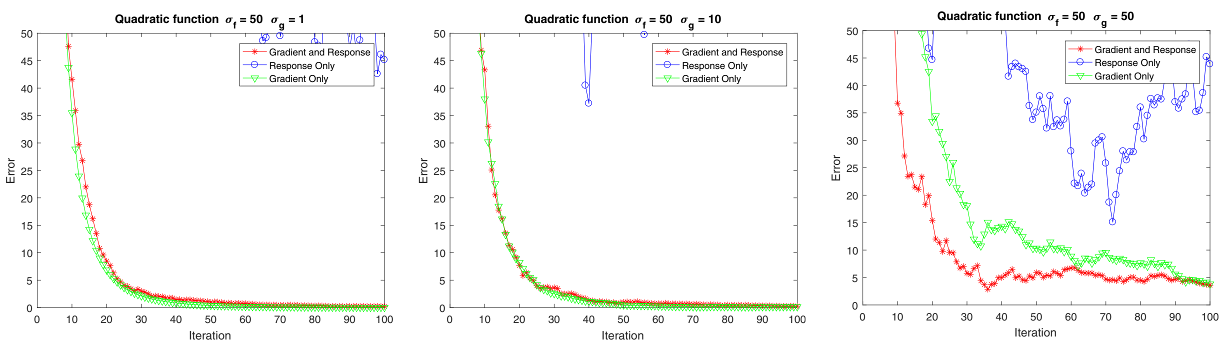


(a) $\sigma_g^2 = 1$.

(b) $\sigma_g^2 = 10$.

(c) $\sigma_g^2 = 50$.

Figure 2: Quadratic function: $\sigma_f^2 = 10$.



(a) $\sigma_g^2 = 1$.

(b) $\sigma_g^2 = 10$.

(c) $\sigma_g^2 = 50$.

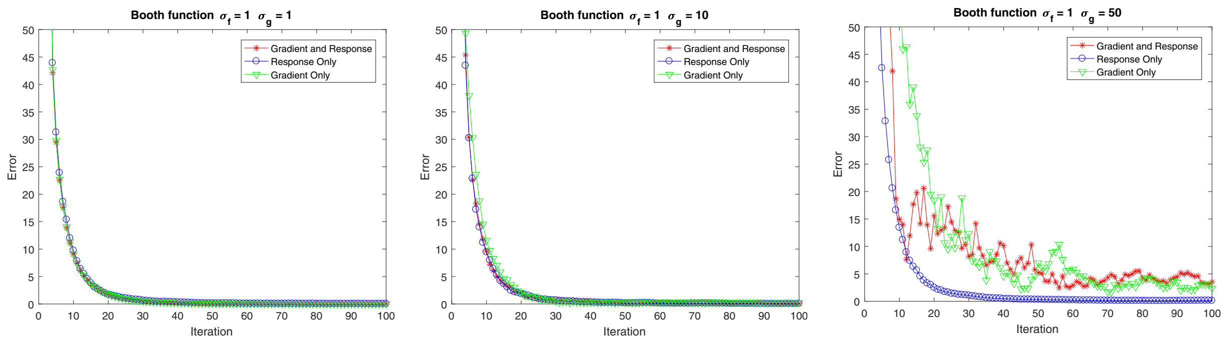
Figure 3: Quadratic function: $\sigma_f^2 = 50$.

in Figures 4 to 6. In Figure 6, design 1 is out of range because of the large error. The results of three designs on Booth function are similar to that of quadratic function: the traditional RSM (Design 2) works well when the response variance is low and becomes unreliable when such variance is large. When only gradient information is utilized (Design 3), the algorithm works relatively well when σ_g is low. When significant noises present in both measurements, the DiGARSM (Design 1) is very robust against such noises, and its performance is comparable to Design 3, which can be seen from Figure 6c.

Remark 3 When gradient and response measurements have the same noise variances, e.g., Figures 2b, 3c, 5b and 6c, design 3 has more stable, reliable performance, suggesting that gradients are more informative than responses under the same noise variance. Therefore, in such cases, practitioners may weight more on gradients in DiGARSM to gain better results.

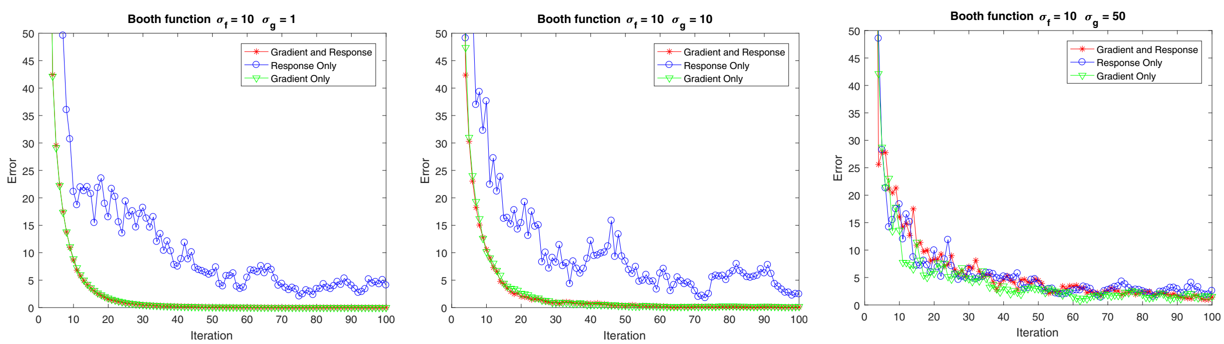
6 CONCLUSION AND FUTURE RESEARCH

In this paper, we proposed DiGARSM, a new metamodel-based optimization method that combines traditional RSM and gradient measurements. In some optimization settings, estimates of the gradient of the objective function w.r.t. the parameters of interest may also be available in addition to responses of the function. Our approach provides a method to combine the response and gradient measurements through $\alpha-$DiGAR model. It was shown by numerical examples that our approach is robust to measurement noise of both information compared to designs that use only single information. We assumed that the objective function has homogeneous variance. Therefore, one possible extension of this work is investigating the performance



(a) $\sigma_g^2 = 1$. (b) $\sigma_g^2 = 10$. (c) $\sigma_g^2 = 50$.

Figure 4: Booth function: $\sigma_f^2 = 1$.



(a) $\sigma_g^2 = 1$. (b) $\sigma_g^2 = 10$. (c) $\sigma_g^2 = 50$.

Figure 5: Booth function: $\sigma_f^2 = 10$.

of DiGARSM on heterogeneous variance. In particular, we expect DiGARSM to have better performance when the variance is proportional to the function value. Another natural extension is to design an algorithm that intelligently chooses the weights in DiGARSM. We expect the algorithm will have better performance if it can adaptively emphasize more on less noisy estimates.

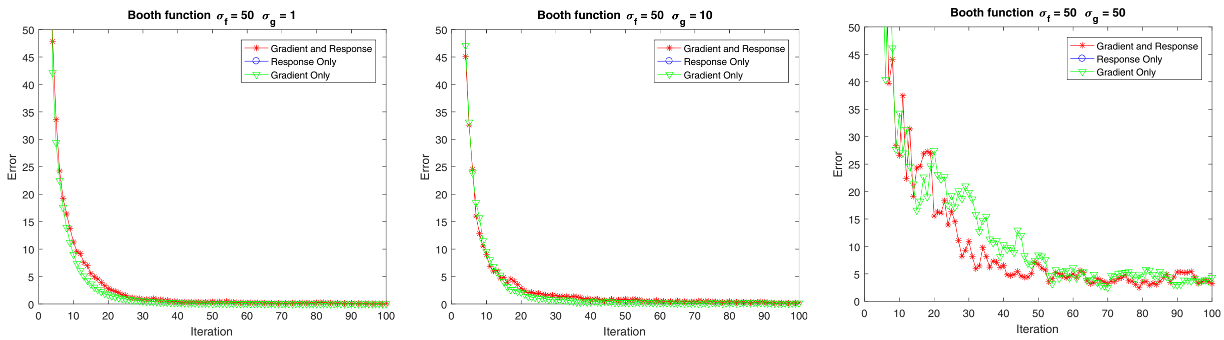
We note again that the usual sequential RSM procedure involves two phases, with the second phase most commonly a single iteration of a higher-order (usually quadratic) fit, from which the optimum is estimated. Such a procedure is necessarily a finite-time procedure involving a stopping time, and thus asymptotic convergence analysis such as employed here would not be appropriate. Future work would consider this setting, and one proposed approach is to use finite-time complexity bounds like the ones introduced in Nemirovski et al. (2009), and summarized in Ghadimi and Lan (2015).

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grants CMMI-1362303 and CMMI-1434419.

REFERENCES

Barton, R. R., and M. Meckesheimer. 2006. "Metamodel-Based Simulation Optimization". Volume 13 of *Handbooks in Operations Research and Management Science*, 535–574. Elsevier.



(a) $\sigma_g^2 = 1$. (b) $\sigma_g^2 = 10$. (c) $\sigma_g^2 = 50$.

Figure 6: Booth function: $\sigma_f^2 = 50$.

Bartz-Beielstein, T., and M. Preuss. 2007. "Experimental Research in Evolutionary Computation". In *Proceedings of the 9th Annual Conference Companion on Genetic and Evolutionary Computation*, edited by D. Thierens, GECCO '07, 3001–3020. New York, NY, USA: ACM.

Box, G. E., and N. R. Draper. 1975. "Robust Designs". *Biometrika* 62(2):347–352.

Box, J., and W. Wilson. 1951. "Central Composites Design". *Journal of the Royal Statistical Society* 1:1–35.

Cahya, S., E. D. Castillo, and J. J. Peterson. 2004. "Computation of Confidence Regions for Optimal Factor Levels in Constrained Response Surface Problems". *Journal of Computational and Graphical Statistics* 13(2):499–518.

Carson, Y., and A. Maria. 1997, 12. "Simulation Optimization: Methods and Applications". In *Proceedings of the 1997 Winter Simulation Conference*, edited by S. Andradottir et al., 118–126. Piscataway, New Jersey: IEEE.

Chang, K.-H., L. J. Hong, and H. Wan. 2007. "Stochastic Trust Region Gradient-Free Method (STRONG)-a New Response-Surface-Based Algorithm in Simulation Optimization". In *Proceedings of the 2007 Winter Simulation Conference*, edited by S. Andradottir et al., 346–354. Piscataway, New Jersey: IEEE.

Chau, M., M. C. Fu, H. Qu, and I. O. Ryzhov. 2014a. "Simulation Optimization: A Tutorial Overview and Recent Developments in Gradient-Based Methods". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk et al., 21–35. Piscataway, New Jersey: IEEE.

Chau, M., H. Qu, and M. C. Fu. 2014b. "A New Hybrid Stochastic Approximation Algorithm". *IFAC Proceedings Volumes* 47(2):241 – 246. 12th IFAC International Workshop on Discrete Event Systems.

Fu, M. C. 1994, Dec. "Optimization via Simulation: A Review". *Annals of Operations Research* 53(1):199–247.

Fu, M. C., and H. Qu. 2014. "Regression Models Augmented with Direct Stochastic Gradient Estimators". *INFORMS Journal on Computing* 26(3):484–499.

Ghadimi, S., and G. Lan. 2015. "Stochastic Approximation Methods and Their Finite-Time Convergence Properties". In *Handbook of Simulation Optimization*, edited by M. C. Fu, Chapter 7, 179–206. Springer.

Glasserman, P. 1991. *Gradient Estimation via Perturbation Analysis*. Springer.

Ho, Y. C., and X. Cao. 1983. "Perturbation Analysis and Optimization of Queueing Networks". *Journal of Optimization Theory and Applications* 40(4):559–582.

Ho, Y.-C., L. Shi, L. Dai, and W.-B. Gong. 1992. "Optimizing Discrete Event Dynamic Systems via the Gradient Surface Method". *Discrete Event Dynamic Systems* 2(2):99–120.

Horng, J.-T., N.-M. Liu, and K.-T. Chiang. 2008. "Investigating the Machinability Evaluation of Hadfield Steel in the Hard Turning with Al2O3/TiC Mixed Ceramic Cool Based on the Response Surface Methodology". *Journal of Materials Processing Technology* 208(1):532 – 541.

Jamil, M., and X.-S. Yang. 2013. "A Literature Survey of Benchmark Functions For Global Optimization Problems". *arXiv preprint arXiv:1308.4008*.

Keyzer, F., J. Kleijnen, E. Mullenders, and A. van Reeken. 1981. "Optimization of Priority Class Queues, with a Computer Center Case Study". *American Journal of Mathematical and Management Sciences* 1(4):341–358.

Kim, W. B., and N. R. Draper. 1994. "Choosing a Design for Straight Line Fits to Two Correlated Responses". *Statistica Sinica* 4(1):275–280.

Kleijnen, J. P. C. 2015. "Response Surface Methodology". In *Handbook of Simulation Optimization*, edited by M. C. Fu, 81–104. New York, NY: Springer New York.

Krafft, O., and M. Schaefer. 1992. "D-optimal Designs for a Multivariate Regression Model". *Journal of Multivariate Analysis* 42(1):130–140.

Law, A. M., W. D. Kelton, and W. D. Kelton. 2007. *Simulation Modeling and Analysis*. McGraw-Hill New York.

Mead, R., and D. J. Pike. 1975. "A Biometrics Invited Paper. A Review of Response Surface Methodology from a Biometric Viewpoint". *Biometrics* 31(4):803–851.

Miró-Quesada, G., and E. Del Castillo. 2004. "An Enhanced Recursive Stopping Rule for Steepest Ascent Searches in Response Surface Methodology". *Communications in Statistics-Simulation and Computation* 33(1):201–228.

Myers, R. H., A. I. Khuri, and W. H. Carter. 1989. "Response Surface Methodology: 1966-1988". *Technometrics* 31(2):137–157.

Nemirovski, A. S., A. Juditsky, G. Lan, and A. Shapiro. 2009. "Robust Stochastic Approximation Approach to Stochastic Programming". *SIAM Journal on Optimization* 19(4):1574–1609.

Plackett, R. L., and J. P. Burman. 1946. "The Design of Optimum Multifactorial Experiments". *Biometrika* 33(4):305–325.

Rubinstein, R. Y., and A. Shapiro. 1993. *Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. John Wiley & Sons Inc.

Wu, S.-M. 1964. "Tool-life Testing by Response Surface Methodology: Part 1". *Journal of Engineering for Industry* 86(2):105–110.

AUTHOR BIOGRAPHIES

YUNCHUAN LI is a PhD student in the Department of Electrical and Computer Engineering and the Institute for Systems Research at the University of Maryland. He received the Bachelor degree in Automation at the University of Electronic Science and Technology of China in 2015. His research interests include simulation optimization, optimization under uncertainty and reinforcement learning. His email address is yli93@terpmail.umd.edu.

MICHAEL C. FU holds the Smith Chair of Management Science in the Robert H. Smith School of Business, with a joint appointment in the Institute for Systems Research, A. James Clark School of Engineering, University of Maryland, College Park. He has a Ph.D. in applied math from Harvard and degrees in math and EECS from MIT. He attended his first WSC in 1988 and served as WSC2011 Program Chair, NSF Operations Research Program Director, *Management Science* Stochastic Models and Simulation Department Editor, and *Operations Research* Simulation Area Editor. He is a Fellow of **INFORMS** and **IEEE**. His e-mail address is mfu@umd.edu, and his Web page is https://www.rhsmith.umd.edu/directory/michael-fu.